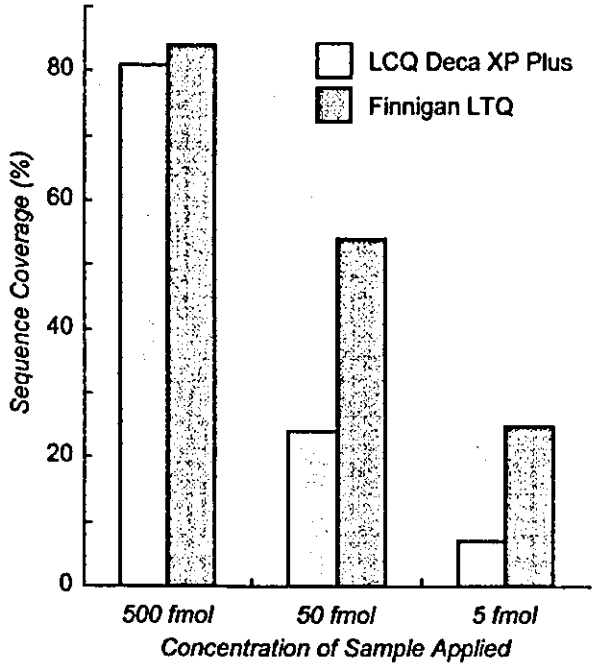

Figure 2. Comparison of sequence coverage of BSA digests (5–500 fmol) by μLC-MS/MS analysis using LCQ Deca XP plus 3-D ion-trap and Finnigan LTQ linear ion-trap MS instruments.

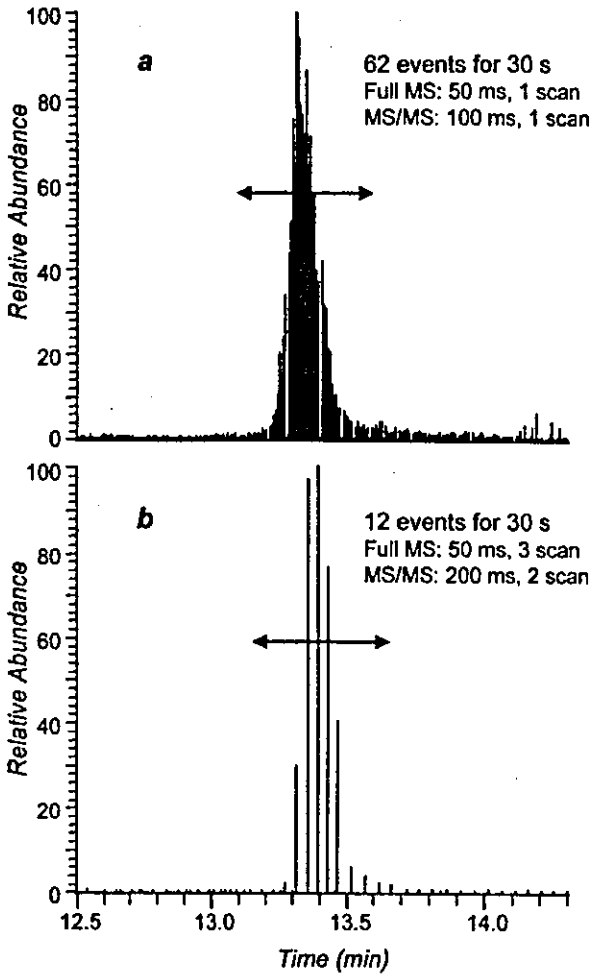

Figure 3. Mass chromatograms at *m/z* 710.0-711.5 by μLC-MS/MS analysis of BSA digests using Finnigan LTQ (a) and LCQ Deca XP plus (b) instruments. A stick in the peak is a single full MS and MS/MS scan of the mass spectrometer, and 62 and 12 events (acquisition of a full MS and a MS/MS in one event) were carried out by the LTQ and LCQ instruments, respectively.

MS was used for comprehensive proteome analysis, several researchers investigated various methods to data-dependently amass MS/MS spectra for a single analysis. These methods include employing a longer analytical time, triple and more MS/MS acquisition against a single full MS spectrum, multiple analyses of the same sample using common conditions or fractionated mass range, etc. While obtaining the MS/MS spectra, the scan speed is a mechanical limitation and a data-dependent scan misses many of the peptide sequences for low abundance peaks that are behind large peaks. Therefore, it is necessary to choose the applications of these techniques for comprehensive proteome analysis of highly complex protein mixtures such as human plasma and whole cell lysates. The drawbacks of clinical proteomics for a large number of human samples are the inability to conduct multiple analyses of the same sample and the longer running time required by the comprehensive LC-MS/MS analysis. Consequently, the performance of the LTQ instrument with a higher scan speed is better than that of the conventional 3-D ITMS instrument, because it enables more informative high-throughput LC-MS/MS analysis for highly complex clinical samples. Therefore, to verify the applicability of LTQ instruments, the human 26S proteasome, which is a highly complex protein mixture consisting of 31 components, was subjected to analysis. As shown in Fig. 1b, equivalent amounts of the digested 26S proteasome sample were analyzed using the LCQ and LTQ instruments under identical μLC conditions. The data-dependent MS/MS acquisition, in which the full MS acquisition is followed by a single MS/MS scan of the most intense precursor ion obtained from the full MS scan, was applied with three microscan full MS (50 ms trapping time) and two microscan MS/MS (200 ms) accumulations for the LCQ, and one microscan of both full MS (50 ms) and MS/MS (100 ms) accumulations for the LTQ instruments. During the 20 min analysis, approximately 450 and 2200 MS/MS spectra were obtained from the 1-D RP μLC-MS/MS analyses using the LCQ and LTQ instruments, respectively. These data were evaluated using a Mascot database search against the Swiss-Prot database, and the search results obtained for the peptide MS/MS assignment were filtered based on the criterion defined as a Mascot peptide score more than 20 and ranked first, described in detail below. Table 1 (see also Supplemen-

Table 1. Protein identification results of human 26S proteasome.

Protein	LCQ Deca XP Plus			Finnigan LTQ		
	Score	Coverage	Peptide	Score	Coverage	Peptide
26S protease regulatory subunits						
PRS4	221	13	3	851	41	16
PRS6	142	9	3	757	48	16
PRS7	348	15	5	1346	55	23
PRS8	299	16	4	1304	58	20
PRSA	377	21	7	1093	53	19
PRSX	154	7	2	715	36	12
Proteasome subunit alpha types						
PSA1	494	36	8	658	48	11
PSA2	261	18	4	569	56	9
PSA3	145	13	3	584	43	11
PSA4	115	9	2	459	33	7
PSA5	158	17	3	483	47	8
PSA6	559	38	8	700	48	11
PSA7	423	35	7	772	57	12
Proteasome subunit beta types						
PSB1	275	34	5	596	53	10
PSB2	256	24	4	456	43	8
PSB3	235	24	3	379	35	5
PSB4	117	9	2	511	48	8
PSB5	528	42	7	670	58	10
PSB6	200	13	3	271	24	5
PSB7	232	14	4	381	33	7
PSB8	ND	ND	ND	180	18	4
26S proteasome non-ATPase regulatory subunits						
PSD1	227	6	3	1636	38	27
PSD2	380	10	6	1386	34	26
PSD3	690	24	10	1445	50	25
PSD4	74	5	1	358	22	6
PSD6	206	12	3	1001	46	18
PSD7	250	15	3	487	42	10
PSD8	ND	ND	ND	228	19	5
PSDB	571	22	8	1445	59	23
PSDC	359	15	5	1147	37	20
PSDD	183	10	4	966	52	17
		Total	130		Total	409

The protein identification data of 31 components consisted of human 26S proteasome including the number of the peptide fragments assigned (Peptide) and the sequence coverage according to these peptides (Coverage). The protein score is calculated by the addition of these peptide scores (Score) in comparison to 1-D μLC-MS/MS analysis using conventional 3-D ion-trap MS (LCQ Deca XP Plus) and new linear ion-trap MS instruments (Finnigan LTQ). ND, not detected.

tary Table A) shows protein identification results including the number of peptide fragments assigned, sequence coverage with these peptides, and the protein score calculated by addition of these peptide scores with respect to the 31 components of the 26S proteasome. In the case of the LTQ instrument, 409 peptide fragments were assigned as components of the 26S proteasome, and this number was approximately three-fold higher than that obtained when the LCQ instrument was used (130 peptide fragments). The individual components of the 26S proteasome were identified by 13.2 and 4.5 peptide fragments, 43.0% and 18.1% sequence coverage, and 821.9 and 292.4 protein scores on an average, using the LTQ and LCQ instruments, respectively. Twenty-nine proteins in 31 components were identified even by the LCQ instruments using our RP μLC system with high-resolution power, as shown in Fig. 1b. However, the peptide fragments of the remaining two components were not observed from the filtered database search results. A number

of peptide fragments belonging to the 26S proteasome (>700 peptides and >60% coverage on an average) were detected by an on-line 2-D SCX/RP μLC-MS/MS experiment with the same amount of digests using the LCQ instrument [10]. These results may simply indicate the difference in the scan speed for the limited, short analytical time between the LCQ and LTQ instruments and not a difference in the sensitivity. Accordingly, LC-MS/MS analysis using LTQ has a three-fold higher efficiency in identification in comparison with a conventional LCQ. This indicates that the LTQ has a superior protein identification capability. Thus, the introduction of LTQ into our μLC-MS/MS system resulted in a highly improved performance, in terms of both sensitivity and protein identification efficiency, for highly complex mixtures.

3.2 Human plasma proteome analysis

The usefulness and applicability of our automated protein profiling system coupled with LTQ for clinical proteomics have been examined by analyzing human plasma samples. In the course of clinical plasma proteome studies, we investigated two types of human plasma – one from healthy donors and the other from donors with lung adenocarcinoma. All plasma samples obtained from the three healthy (H-N, H-I, and H-S) and two adenocarcinoma (AC88 and AC94) cases were digested in a solution with trypsin after removing their HSA and IgG contents. The resultant peptide mixtures were diluted for the μLC-MS/MS analysis. Further, equivalent mixture samples from the three healthy donors (H-NIS; H-N:H-I:H-S, 1:1:1 in volume) and the two adenocarcinoma donors (AC8894; AC88:AC94, 1:1 in volume) were also prepared as average samples for each case. The individual and mixture samples (total seven samples) were analyzed by 1-D RP μLC-MS/MS analysis under analytical conditions for 90 min. The LTQ MS data was acquired by the double MS/MS method, in which the full MS acquisition is followed by two MS/MS scans of the two most intense precursor ions from the full MS scan. This is done to improve the protein identification results by the database search. Additionally, our established on-line 2-D SCX/RP μLC-MS/MS system using LTQ instead of the conventional LCQ instrument was also used to analyze the mixture samples (H-NIS and AC8894) along with an additional 1-D RP analysis. In the 2-D μLC-MS/MS analysis, six SCX separation runs were automatically carried out and analyzed by the RP μLC-MS/MS analysis under the same analytical conditions as described earlier. The total operation time for both 1-D and 2-D μLC-MS/MS analyses of a sample was within 11 h [10]. Figure 4 shows the base-peak chromatograms from 1-D (g) and 2-D

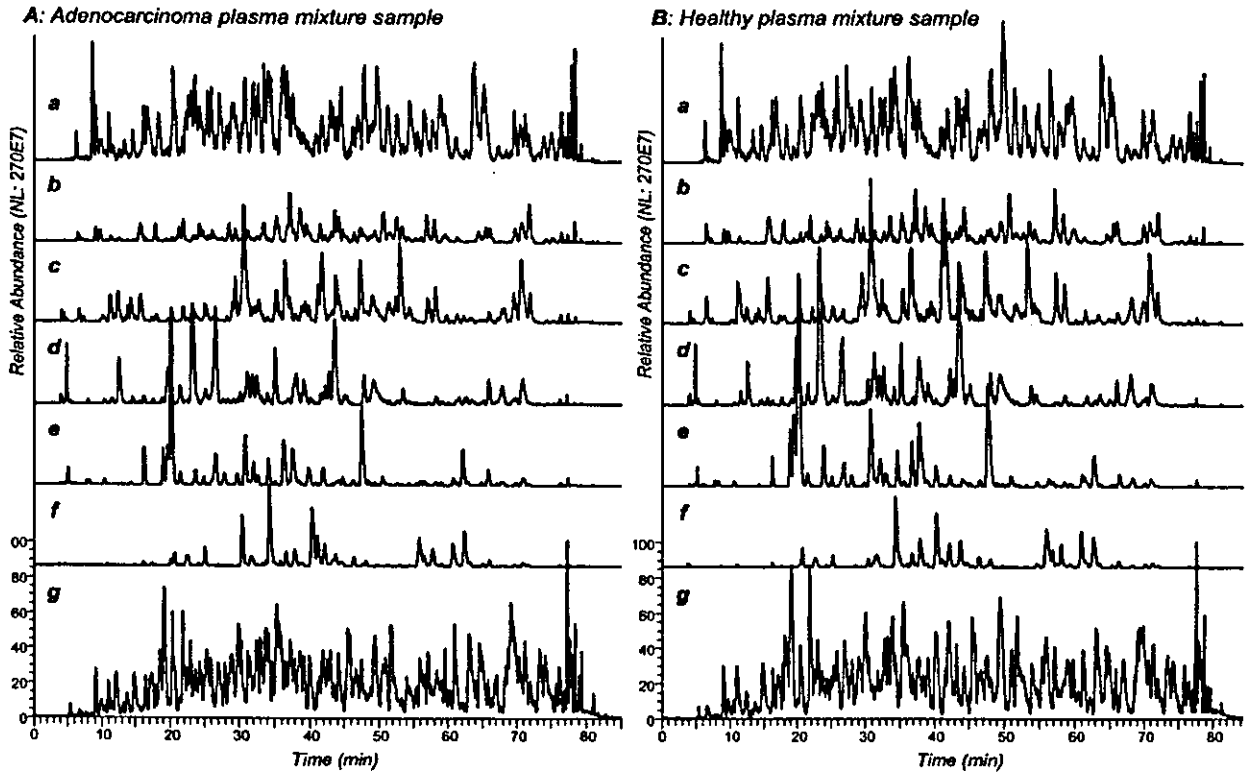

Figure 4. Base-peak chromatograms of the digested human plasma samples using 2-D SCX/RP (a–f) and 1-D RP μLC-MS/MS analyses (g). A, mixture of plasma digests of the healthy group; B, mixture of plasma digests of the adenocarcinoma group; a, 25 mM; b, 50 mM; c, 100 mM; d, 150 mM; e, 200 mM; f, 500 mM salt concentration SCX fractions for 2-D μLC-MS/MS analysis.

μLC-MS/MS analyses (a–f) of approximately 2 μg AID-HP tryptic digests corresponding to 0.4 μL original blood plasma sample. The 1-D RP μLC-MS/MS analysis provided approximately 10 000 MS/MS spectra for each sample, and the resultant data were evaluated using a Mascot database search against *H. sapien* subsets of the sequences in the Swiss-Prot database.

In order to achieve statistical confidence levels in identification of proteins from highly complex mixture samples, we investigated the thresholds as filters to extract data for the Mascot peptide score. We used the datasets, obtained from the 1-D μLC-MS/MS analysis with LTQ, of the digested 26S proteasome sample within our search tolerances. Since it is possible to identify several proteins from a single MS/MS spectrum based on the hit sequence varieties, the most significant peptide sequence that was ranked first (marked with bold red in the Mascot search results), which had the highest score among the hit varieties, was extracted from the entire datasets to prevent erroneous identifications and redundancy. The resulting peptide assignments were sorted according to their Mascot peptide score and intergraded into protein identification. Figure 5 (see also Supplementary Table A) shows the relationship between the peptide score ranges and either the number of peptide fragments assigned as 26S proteasome or other proteins identified erroneously. Although peptides with a score less than 50 were assigned to both the 26S proteasome and to the other proteins, all peptide fragments with a score of more than 50 were confidently identified as belonging to the 26S proteasome. For thresholds of peptide scores higher than 20, 30, and 40, the statistical identification confidence levels of the Mascot database search results were 70%, 90%, and 97%, respectively. Accordingly, we tentatively set peptide scores more than 30 and ranked first as the criterion for a broad protein identification index in order to integrate the datasets of the μLC-MS/MS analysis of plasma samples. Furthermore, to extract plasma proteins with a higher confidence level, we finally tried to apply the Swiss-Prot dataset (667 proteins) in a non-redundant list of 1175 distinct proteins that Anderson *et al.* have recently developed by combining four separate sources of human plasma proteome [3, 13–16].

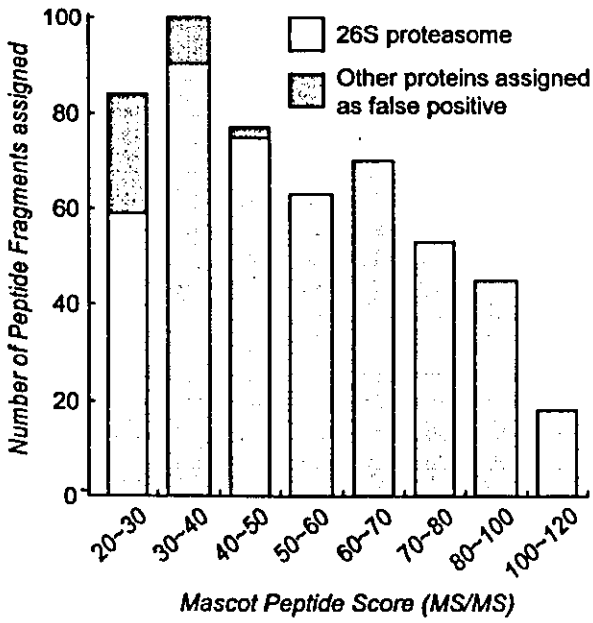

Figure 5. Mascot database search results of 1-D μLC-MS/MS analysis of the digested 26S proteasome. See also Supplementary Table A.

The Mascot database search results on plasma proteome analysis yielded data extracted under the thresholds of peptide scores higher than 30 and ranked first; an average of 108 proteins were detected from each 1-D μLC-MS/MS analysis. From the 2-D μLC-MS/MS analysis of the mixture sample of two groups, an average of 249 proteins was assigned as plasma proteome candidates. (Supplementary Table B). Additionally, entire datasets of these Mascot search results were integrated and processed with the data extraction. The results indicated that a total of 506 different proteins were listed, and 180 proteins were detected as common proteins from both groups (Fig. 6, in parentheses). Furthermore, plasma proteins with a high confidence level were extracted from these datasets using the Swiss-Prot dataset of the 667 plasma proteins reported. Figure 6 shows the diagrammatic representation of proteins found in the two groups comprising healthy individuals and adenocarcinoma patients, and the numbers are concordant with the proteins annotated as plasma proteins in 667 Swiss-Prot datasets. The results indicated that 84 and 85 proteins were extracted from the healthy and adenocarcinoma groups, respectively, and 69 proteins were common. In addition, 16 proteins were

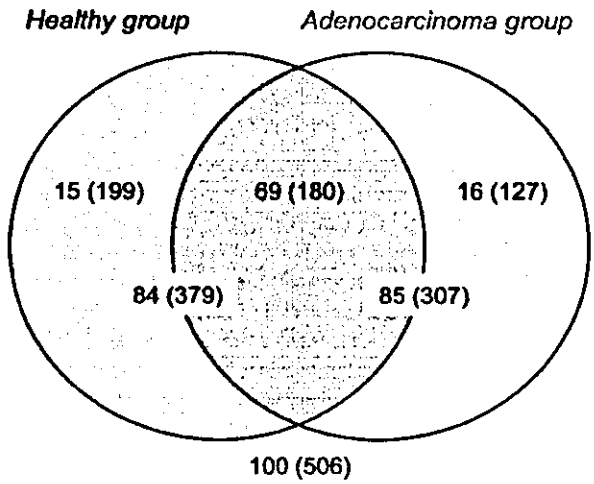

Figure 6. Diagrammatic representation of the proteins detected in the groups of healthy individuals and adenocarcinoma patients by 1-D and 2-D μLC-MS/MS analyses of their plasma samples. Numbers that belong to the peptide score thresholds higher than 30 served as the criteria for protein extraction (expressed in parentheses), and the numbers are concordant with the plasma proteins reported in the 667 Swiss-Prot datasets.

Table 2. Selected specific protein list of the healthy and lung adenocarcinoma groups in human blood plasma.

Accession No.	Protein name	Adenocarcinoma plasma samples					Healthy plasma samples					
		2-D analysis	1-D μLC-MS/MS analysis				2-D analysis	1-D μLC-MS/MS analysis				
		AC8894	AC88	AC94	AC8894-1	AC8894-2	H-NIS	H-N	H-I	H-S	H-NIS-1	H-NIS-2
P02649	Apolipoprotein E	A	C	A	A	A			C			
Q14624	Inter-alpha-trypsin inhibitor heavy chain H4	A	C	B	C		C		C	C	C	C
P04196	Histidine-rich glycoprotein	A										
P00748	Coagulation factor XII	B										
P00488	Coagulation factor XIII A chain	B					C					
P02570	Actin, cytoplasmic 1	C					A	A	C		C	
P27169	Serum paraoxonase/arylesterase 1						A					
P29312	14-3-3 protein zeta/delta								A			
P54108	Cysteine-rich secretory protein-3	C					B					
O14791	Apolipoprotein L1						B					
P02751	Fibronectin						B					
P06396	Gelsolin, plasma						B					

Two adenocarcinoma plasma samples (AC88 and AC94) and three healthy plasma samples (H-N, H-I, and H-S) were analyzed by 1-D μLC-MS/MS analysis. The mixture samples separated into both groups (AC8894 and H-NIS) were analyzed twice by 1-D and once by 2-D μLC-MS/MS analysis (2-D analysis). A–C indicated the presence of peptide fragment(s) assigned to the listed protein. A, mascot peptide score higher than 50; B, score 40 to 50; C, score 20 to 40. See also Supplementary Table B.

detected as specific significant proteins of the adenocarcinoma group, and 15 proteins were not detected in the adenocarcinoma group. Table 2 (see also Supplementary Table B) shows a list of the specific proteins assigned by peptide fragment(s) with scores higher than 40. These specific proteins detected from only one group might be candidate biomarkers of lung adenocarcinoma in human blood plasma. However, further statistical verification of our results through data accumulation of more disease plasma samples and the investigation concerning the reproducibility of protein identifications for each sample are necessary. Additionally, validation of these protein identifications by several biochemical approaches would be required. In the present study, we could indicate that several significant protein candidates in the plasma proteome are possibly associated with the pathological differences in lung adenocarcinoma. Functions of specific proteins and their correlations with adenocarcinoma along with the other proteins are not listed in this paper and will be reported elsewhere. These experimental achievements suggest that our automated 1-D and 2-D μLC-MS/MS protein profiling systems, in which the LTQ was incorporated, are powerful in identifying low-abundance proteins of great clinical importance, because these molecules directly report the occurrence and progress of various diseases.

4 Concluding remarks

In the course of the Human Plasma Organization Plasma Proteome Project, several research groups have prepared contrast reference specimens of human plasma using various technology platforms [17]. It is extremely important to catalog the plasma proteome as a protein database for clinical plasma proteomics. Applied technology platforms are very powerful, particularly for comprehensive broad protein identification of highly complex samples such as blood plasma. However, it seems difficult to stably and reproducibly apply these to routine clinical investigations that require a large number of proteome analysis runs for a large number of human samples. We have recently established a fully automated, high-throughput 2-D SCX/RP μLC-MS/MS protein profiling system, which can perform large-scale analysis in clinical proteomics [10]. In this study, the LTQ, which is superior to a conventional 3-D ITMS instrument in sensitivity and scan speed, was utilized in our high-throughput system, and it was evaluated by analyzing BSA and human 26S proteasome. Furthermore, the system was applied to plasma proteome analysis in a few cases of both healthy individuals and lung adenocarcinoma patients. The results confirmed that a 10-fold increase in terms of sensitivity was achieved in our system using the LTQ instrument for protein

identification. Further, in comparison with the conventional 3-D ITMS instrument, a three-fold higher number of peptide fragments was identified as belonging to the 26S proteasome, indicating significant improvement in resolution for the analytical time point. Additionally, approximately 250 and 100 different proteins were detected, based on the investigation criterion for a 90% confidence level of protein identification, from only 0.4 μL human plasma using 2-D and 1-D μLC-MS/MS analyses, respectively. The entire operation was automatically carried out within 11 h for both single 1-D and 2-D μLC-MS/MS analyses. From the protein identification datasets of both healthy and adenocarcinoma plasma samples, several disease-specific proteins were found in the human plasma based on the plasma proteome database reported earlier. Consequently, it was demonstrated that our μLC-MS/MS protein profiling system is feasible for large-scale analyses such as clinical plasma proteomics studies. Although plasma proteome analysis for clinical application still remains a great challenge due to the wide dynamic range of protein abundance, we shall continue further technological development of the large-scale proteome analysis based on the high-throughput μLC-MS/MS system reported in this paper. Such high-throughput and large-scale analysis of human plasma would lead to the discovery of new disease-associated protein markers with high sensitivity and high specificity in early disease detection and diagnosis, and this would revolutionize current therapeutics.

The authors gratefully acknowledge the technical assistance of Ms. Hisae Anyoji and Noriko Araki of Medical Proteoscope, Co. Inc., and medical doctors of the Department of Surgery, Tokyo Medical University, as well as the encouragement and support of AMR Inc., Tokyo, Japan. We are also deeply indebted to Drs. Hiroshi Matsumoto and Masayuki Kubota of Thermo Electron Co., Kanagawa, Japan for their assistance.

5 References

[1] Liotta, L. A., Ferrari, M., Petricoin, E., *Nature* 2003, *425*, 905.

[2] Anderson, N. L., Anderson, N. G., *Mol. Cell. Proteomics* 2002, *1*, 845–867.

[3] Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 311–326.

[4] Washburn, M. P., Wolters, D., Yates, J. R. III, *Nat. Biotechnol.* 2001, *19*, 242–247.

[5] Wolters, D. A., Washburn, M. P., Yates, J. R., III, *Anal. Chem.* 2001, *73*, 5683–5690.

[6] Shen, Y., Jacobs, J. M., Camp, D. G., II, Fang, R. *et al.*, *Anal. Chem.* 2004, *76*, 1134–1144.

[7] Kawakami, T., Nagata, T., Muraguchi, A., Nishimura, T., *Electrophoresis* 2000, *21*, 1846–1852.

[8] Kawakami, T., Nagata, T., Muraguchi, A., Nishimura, T., *J. Chromatogr. B* 2003, *87*, 223–229.

[9] Fujii, K., Nakano, T., Kawamura, T., Usui, F. *et al.*, *J. Proteome Res.* 2004, *3*, 712–718.

[10] Fujii, K., Nakano, T., Hike, H., Usui, F. *et al.*, *J. Chromatogr. A* 2004, *1057*, 107–113.

[11] Tojo, H., *J. Chromatogr. A* 2004, *1056*, 223–228.

[12] http://www.matrixscience.com

[13] Pieper, R., Su, Q., Gatlin, C. L., Huang, S. T. *et al.*, *Proteomics* 2003, *3*, 422–432.

[14] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 947–955.

[15] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J. *et al.*, *Mol. Cell. Proteomics* 2003, *2*, 1096–1103.

[16] Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S. *et al.*, *Proteomics* 2003, *3*, 1345–1364.

[17] Omenn, G. S., *Proteomics* 2004, *4*, 1235–1240.

DOI 10.1002/pmic.200401047

REGULAR ARTICLE

Protein identification from product ion spectra of peptides validated by correlation between measured and predicted elution times in liquid chromatography/mass spectrometry

Takao Kawakami[1, 2], *Keita Tateishi*[2], *Yuko Yamano*[3], *Takashi Ishikawa*[4], *Kazuyuki Kuroki*[3], *and Toshihide Nishimura*[1, 2]

[1] Clinical Proteome Center, Tokyo Medical University, Tokyo, Japan
[2] Medical ProteoScope, Tokyo, Japan
[3] Cancer Research Institute, Kanazawa University, Ishikawa, Japan
[4] Department of Internal Medicine, Graduate School of Medicine, University of Tokyo, Tokyo, Japan

Received: May 14, 2004
Accepted: September 23, 2004

Reversed-phase liquid chromatography (LC) directly coupled with electrospray-tandem mass spectrometry (MS/MS) is a successful choice to obtain a large number of product ion spectra from a complex peptide mixture. We describe a search validation program, ScoreRidge, developed for analysis of LC-MS/MS data. The program validates peptide assignments to product ion spectra resulting from usual probability-based searches against primary structure databases. The validation is based only on correlation between the measured LC elution time of each peptide and the deduced elution time from the amino acid sequence assigned to product ion spectra obtained from the MS/MS analysis of the peptide. Sufficient numbers of probable assignments gave a highly correlative curve. Any peptide assignments within a certain tolerance from the correlation curve were accepted for the following arrangement step to list identified proteins. Using this data validation program, host protein candidates responsible for interaction with human hepatitis B virus core protein were identified from a partially purified protein mixture. The present simple and practical program complements protein identification from usual product ion search algorithms and reduces manual interpretation of the search result data. It will lead to more explicit protein identification from complex peptide mixtures such as whole proteome digests from tissue samples.

Keywords: Elution time prediction / Liquid chromatography / Mass spectrometry / Product ion search / Protein identification

1 Introduction

Protein identification is a critical step in most biological research and especially in proteomics [1]. MS has been a powerful tool for this purpose, mainly because its output (mass spectra) directly links sample protein to sequence. Peptide product ion searches are one of the commonly used methods for mass spectrometric protein identification. Usually, protein mixtures or purified proteins are proteolytically digested in solution or in gel matrix. The resulting peptide mixtures are analyzed by MS/MS instruments for mass-to-charge ratios of precursor ions from respective peptides and those of product ions generated by CID of the precursor ions. Data sets of the product ion spectra are subsequently searched against primary structure databases using search engines to identify the protein(s) contained in the analytes. MS instruments, such as the ion-trap, are routinely applied in 1-D and 2-D LC directly coupled with nanoelectrospray ionization (nESI)-MS/MS in order to generate product ion spectra from

Correspondence: Dr. Takao Kawakami, Clinical Proteome Center, Tokyo Medical University, Shinjuku Sumitomo Building 17, 2-6-1 Nishi-shinjuku, Shinjuku-ku, Tokyo 163-0217, Japan
E-mail: takao30@tokyo-med.ac.jp
Fax: +81-3-5321-6624

Abbreviations: **HBV**, hepatitis B virus; **nESI**, nanoelectrospray ionization; **PSEC**, probability score-based elution correlation

peptides eluted continuously from the LC column [2, 3]. Instrumental controls for CID of automatically selected peptide ions are introduced to efficiently acquire product ion spectra from complex peptide mixtures. In the automated data acquisition mode, thousands of product ion spectra are usually obtained *per* run when analyzing peptide mixtures such as whole proteome digests from tissues.

The data sets often include considerable poor spectra. These are caused by several factors including insufficient amounts of precursor ions injected into the MS collision cell, CID of nonpeptide precursor ions and the properties of peptide ions on proton distributions along the amino acid sequences. The data sets also include product ion spectra without any significant hits in spite of sufficient intensity and number of product ions in the spectra, partially due to PTMs, nonspecific digestions and no corresponding sequence in the databases. Moreover, lower mass accuracy of instruments, such as in the case of ion trap MS, often result in several hit candidates *per* product ion spectrum. These elements increase both false positive and false negative peptide assignments. It complicates interpretation of final protein identification lists resulting from the arrangement of each peptide assignment to a protein structure.

From the report from Meek [4], it is known that absolute peptide elution times in RP LC are predictable with accuracy from the physicochemical properties of the peptides. Many investigators have reported prediction formulas under defined conditions of peptide separation. In most cases, the formulas contain the sum of retention coefficients defined for respective amino acid residues composing a peptide structure and a term to correct for the dead volume of the LC elution. Recently, Petritis *et al.* [5] introduced artificial neural networks in order to predict the RP LC elution times of peptides in LC-MS/MS analysis. In the present report, we demonstrate a ScoreRidge program developed to complement conventional product ion searches. ScoreRidge is based on correlation between observed LC elution times and calculated elution times. Consequently, fine adjustments of the calculation parameters to correct for fluctuations between analyses were minimized. Peptide assignments within a given tolerance of the correlation were accepted to arrange to protein sequences in the databases. The program was applied to identify host protein candidates binding to hepatitis B virus (HBV) core protein, one of the key components in the life cycle of HBV.

2 Materials and methods

2.1 Materials

A protein complex of *Saccharomyces cerevisiae* 20S proteasome was purchased from AFFINITI Research Products (Exeter, UK). This sample was used for the program evaluation. Three phosphopeptides, DRVpYIHPF, LSRQLpSS GVSEC and KRELVEPLpTPSGEAPNQALLR, and their nonphosphorylrated forms were synthesized by the American Peptide Company (Sunnyvale, CA, USA). LC grade ACN and formic acid were obtained from Wako Pure Chemical Industries (Osaka, Japan), and the water used was Milli-Q analytical grade (Millipore, Bedford, MA, USA). Sequencing grade trypsin was from Promega (Madison, WI, USA).

2.2 Partial purification of human HBV-binding host proteins

Glutathione-S-transferase (GST)-HBV core fusion proteins were expressed in *Escherichia coli*. The fusion proteins were prepared with glutathione beads as described [6]. HepG2 cells were lysed according to [7], and the cell lysate was pretreated with a glutathione-sepharose column in order to remove host proteins which bound nonspecifically to glutathione molecules. The lysate was then incubated with the bead-immobilized GST-HBV core protein. The beads were washed with lysis buffer and incubated with SDS-sample buffer to dissociate and denature the bound proteins. The proteins were subjected to 12.5% SDS-PAGE. Protein bands separated on the SDS-gel were stained with a SYPRO Ruby fluorescence dye (Genomic Solutions, Chelmsford, MA, USA). A protein band with M_r 33,000 (p33) was subjected to digestion (see Section 2.3).

2.3 Protein digestion

The 20S proteasome complex was dissolved in a solvent of 25% ACN and 25 mM ammonium bicarbonate (NH_4HCO_3), and incubated at 37°C for 30 min. After *S*-alkylation of cysteine residues with iodoacetamide, the proteasome protein subunits were digested with trypsin at 37°C for 16 h. The p33 band was excised from the gel plate and subjected to an in-gel tryptic digestion process according to the method of Shevchenko *et al.* [8]. Peptide mixtures resulting from these treatments were dried under vacuum and stored at −20°C until use.

2.4 LC-MS/MS

RP peptide separation was performed with a MAGIC2002 LC system (Michrom BioResources, Auburn, CA) containing a MAGIC C18 capillary LC column (0.2 mm id, 50 mm length, 3 μm particle size, and 200 Å pore size; Michrom BioResources). Mobile phase A consisted of formic acid, ACN and water at a volume ratio of 0.1:2:98, and mobile phase B consisted of formic acid, ACN and water at a volume ratio of 0.1:90:10. The initial flow of 90 μL/min was split by a MAGIC Splitter (Michrom BioResources) to approximately 1 μL/min. The peptide mixtures were dissolved in a solvent, consisting of TFA, ACN and water at a volume ratio of 0.1:2:98 and injected onto a Peptide CapTrap column (0.5 mm id, 2.0 mm length, bed volume 0.5 μL; Michrom BioResources) equilibrated with mobile phase A. The peptides concentrated and purified on the trap column were injected

onto the C18 capillary LC column by valve switching. The peptides were eluted at a rate of 1 μL/min on the linear gradient with various rates of mobile phase B, mainly depending on the sample complexity injected. The LC effluent was directly interfaced with an electrospray ion source on an LCQ-Deca XP ion trap mass spectrometer (Thermo Electron, San Jose, CA, USA) with a FortisTip capillary needle (20 μm id, top teflon-coated; AMR, Tokyo, Japan) for microflow LC-nESI-MS/MS analysis. nESI used no sheath gas and the experimental parameters included a voltage of 1.7 kV; a tube lens offset of 15 V; a capillary voltage of 37 V; and a capillary temperature of 250°C. Peptides were analyzed in a fully automated fashion by the data-dependent scanning methods of a consecutive full scan (*m/z* 400–2000) and product ion scan of the major precursor peptide ions. The product ion scan was performed with an intensity threshold of 5×10^4, 35% normalized collision energy, 3.0 amu isolation mass width, and dynamic exclusion for 3 min. nESI-MS/MS operation and data acquisition were carried out on an Xcalibur version 1.4 system controller (Thermo Electron).

2.5 Data base searching and the validation processes

2.5.1 Data conversion

Mass chromatogram files were generated for every LC-MS/MS analyses. Product ion spectra in the chromatograms were converted to peak list files in a DTA format using an LCQ_DTA.EXE utility program (Thermo Electron). Each peak list file consisted of the following numerical values: precursor mass + proton mass (MH), charge number of the precursor ion, *m/z* of product ions and ion intensities of the product ions. The file name contains the scan number of the original product ion spectrum. Criteria of product ion spectra for data conversion were: a minimum number of product ions in a spectrum of 10 and a minimum total ion counting threshold of 0. Each product ion spectra was converted automatically into either one peak list file when the precursor ion was singly-charged or two peak list files each when the precursor ion was doubly- or triply-charged.

2.5.2 Data base search for peptide assignments

In order to assign amino acid sequences, the peak lists (see Section 2.5.1) were searched using the MOWSE [9] scoring algorithm, MASCOT [10, 11], version 1.9.0, against the non-redundant proteome databases of Swiss-Prot, TrEMBL and Ensembl entries (http://www.ebi.ac.uk/proteome; updated on 13 April 2004). The protein sequence sets for *S. cerevisiae* (6231 entries), and for *Homo sapiens* (28, 814 entries), were separately down-loaded in a FASTA format and used for the search. Searches were done with the following criteria: tryptic digestion (hydrolysis of the peptide bonds following lysine and arginine residues); fixed modification of cysteine (*S*-carboxyamidomethylation, +57.0); variable modifications of methionine (oxidation, +16.0) and protein amino-terminal (*N*-acetylation, +42.0); one missed cleavage; a peptide mass tolerance of ± 2 amu; a product ion mass tolerance of ± 0.8 amu. Top-scored peptide assignments to respective peak lists were considered for ScoreRidge processing as shown in Section 2.5.3.

2.5.3 Correlation equations based on probability scores

Basically ScoreRidge was programmed and operated on Perl scripts. Peptide elution times were calculated using parameters in a PepStat utility program (Thermo Electron) with modifications and additions on retention coefficient values of modified amino acid residues and *N*-acetylation (Table 1). Peptide assignments, accompanied with the MASCOT probability scores of peptide matching, MS scan number and calculated elution values, were accumulated in the order of probability score heights by a given interval. These accumulated data sets were used to lead correlation equations between the scan numbers and calculated elution values and to calculate the correlation coefficients (r). With accumulation of the peptide assignments, the r-values generally remained almost constant before markedly decreasing. In the constant r-value range, it was expected that a sufficient number of

Table 1. Retention coefficients for amino acid residues

Residue	Retention coefficient
Alanine	2.0
Arginine	−0.6
Asparagine	−0.6
Aspartic acid	0.2
Cysteine	2.6
Glutamic acid	1.1
Glutamine	0.0
Glycine	−0.2
Histidine	−2.1
Isoleucine	7.4
Leucine	8.1
Lysine	−2.1
Methionine	5.5
Phenylalanine	8.1
Proline	2.0
Serine	−0.2
Threonine	0.6
Tryptophane	8.8
Tyrosine	4.5
Valine	5.0
α-Amino	−6.9
α-Carboxyl	−0.8
N-Acetyl	9.1[a]
Carboxyamidomethylcysteine	0.0[a]
Phosphoserine	1.8[a]
Phosphothreonine	2.6[a]
Phosphotyrosine	6.5[a]

a) Determined using the present LC-MS/MS conditions

probable peptide assignments would be accumulated for higher correlation. The correlation equations in the range were defined as probability score-based elution correlation (PSEC) equations. In order to find the distances of the calculated elution values against the PSEC equations of the respective peptide assignments, the calculated elution values were subtracted from the elution values estimated by the PSEC equations. Peptide assignments within a given prediction error value were selected for the arrangement to list protein entries identified.

2.5.4 Arrangement of peptide assignments

Peptide assignments passing the above validation processes (see Section 2.5.3) were arranged again using MASCOT. Protein entries with total matching scores exceeding the probable identification threshold were considered for interpretation of identification.

3 Results and discussion

3.1 Product ion search complemented by peptide elution times

LC-nESI-MS/MS is a successful choice to efficiently obtain a large number of peptide product ion spectra from complex peptide mixtures. However, all product ion spectra are not correctly assigned to amino acid sequences in the sequence databases. Peptide retention on the RP LC is closely related to the hydrophobicity of the peptide molecules, and the LC elution times are important values for confirmation of peptide sequence assignments to the product ion spectra. Whereas peptide elution times in RP LC are predictable with a certain accuracy due to the physicochemical parameters deduced from the peptide sequences, fluctuation of peptide elution conditions in each analysis has made it difficult to steadily use a single prediction equation. This is especially true when using micro- or nano-flow LC developed for highly sensitive analysis. The fluctuations include linear change of elution dead volume and nonlinear instability of gradient slope. They are caused by several factors including lot to lot variation of the LC columns, decline of the quality of the column and LC pump due to their continuous use, and partial stop-up of solvent tubes. For more practical elution prediction, we introduced the PSEC equations.

An MS/MS ions search mode in MASCOTwas a basis for the present product ion search. It is one of the commonly used programs with the highest percentage of correct protein identification [12]. Essentially, ScoreRidge is able to complement other search engines such as Sequest [13, 14], Spectrum Mill [15], in expressing the probability of each peptide assignment. These search engines usually include two functions: assignment of peptide sequences to product ion spectra and arrangement of the peptide assignments onto protein sequences in the primary structure databases. These functions were integrated separately into the present data processing.

3.2 Retention coefficients for modified amino acid residues

A tryptic digest of the 20S proteasome was analyzed with LC-MS/MS. A total of 196 product ion spectra were manually interpreted with the aid of MASCOT. Of them, 154 spectra were interpreted as unmodified tryptic peptides, 35 spectra were from peptides containing one or two *S*-carboxyamidomethylcysteine residue(s), and the remaining seven spectra were from *N*-acetylated peptides. The 154 product ion spectra from unmodified peptides were used as a standard dataset for optimum peptide elution prediction. PepStat gave the highest r-value (0.924) between the scan numbers of the 154 peak lists and the elution values calculated from the amino acid sequences. Scan numbers of the respective peak lists were used as a substitute of the measured retention times of precursor peptides, because both values were correlative with r-values higher than 0.99 at almost regular time intervals of the product ion scan. ScoreRidge was simply operated owing to the substitution. Calculation by PepStat was done without any corrections of output values. Consequently, ScoreRidge uses correlations between scan numbers and calculated elution values and does not consider coincidence between both values.

Table 1 shows retention coefficients of unmodified and modified amino acid residues. The height of these values indicates the degree of contribution of the respective amino acid residues toward the delay in peptide elution. *S*-Carboxyamidomethylation is a commonly used chemical modification to block thiol groups of cysteine residues prior to proteolytic digestion or chemical cleavage of proteins [16]. *N*-Acetylation caused protein *N*-terminal blockage and is frequently observed under physiological conditions [17]. Retention coefficients for these modifications were found by optimum fitting of 35 peptides containing *S*-carboxyamidomethylcysteine residue(s) and seven *N*-acetylated peptides onto the correlation curve from the 154 unmodified peptides.

Figure 1 shows a comparison of peptide elution of three phosphopeptides and their nonphosphorylated forms. A mixture consisting of six peptides was analyzed by LC-MS/MS. Each chromatogram plots the total intensity of both singly- and doubly-charged ions generated from each peptide. The three phosphopeptides contain phosphoserine, phosphothreonine and phosphotyrosine, respectively. With the LC-MS/MS system and solvent compositions used, phosphorylation of peptides resulted in a delay in elution. Retention coefficients of these phospho-amino acids were found by addition of 2.0 units to those of the corresponding unmodified residues in order to fit on the above correlation curve (Table 1).

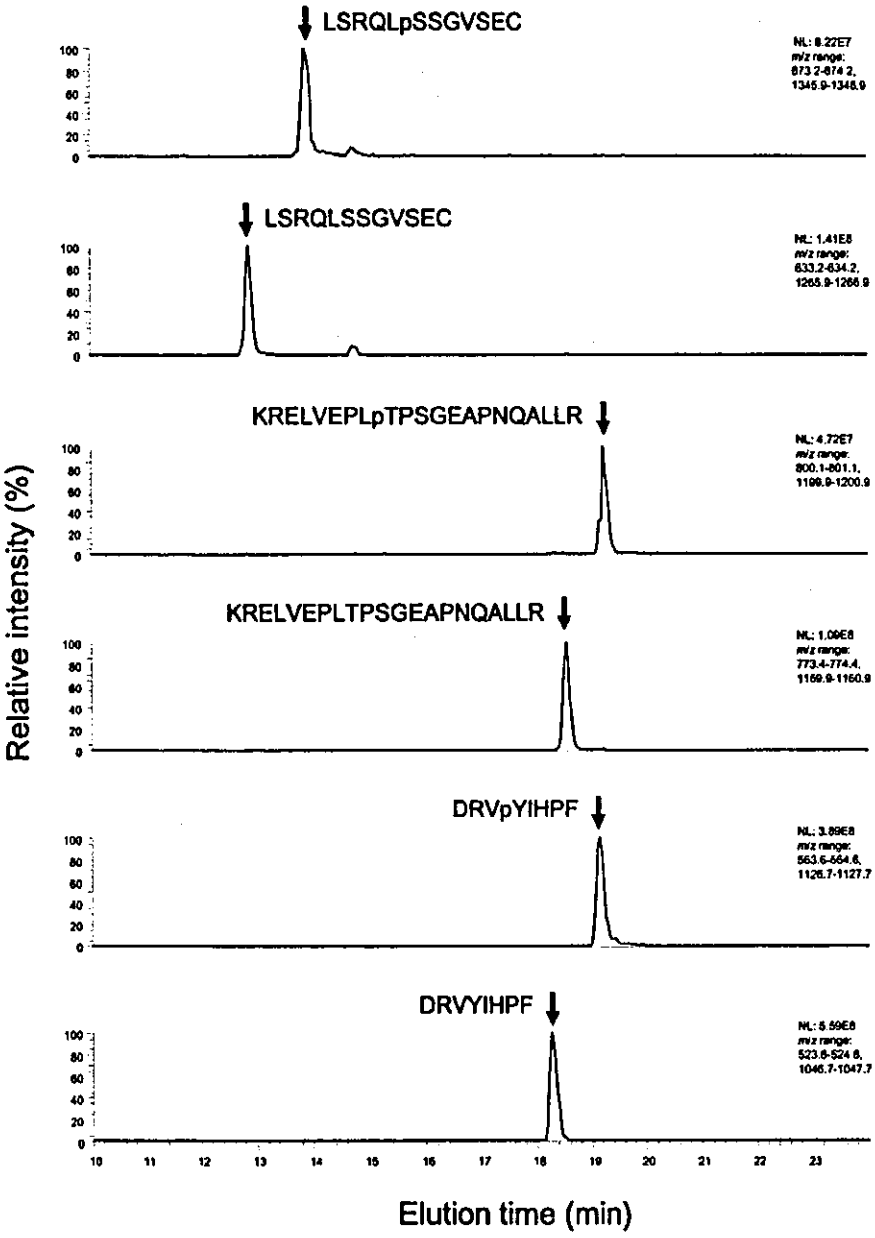

Figure 1. Mass chromatograms of phosphopeptides and their nonphosphorylated forms.

3.3 Data analysis

Using LC-MS/MS data from the 20S proteasome digest, ScoreRidge was simulated as follows. The LC elution conditions were set on the linear gradient with 5% to 45% mobile phase B in 40 min. The total product ion spectra (1188) was obtained in a single analytical run. These product ion spectra were converted to 2134 peak lists as described in section 2.5.1. Obtained ion peak lists were searched for assignment to tryptic peptide fragments in the yeast proteome database. Of the 2134 peak lists, 1837 were assigned to peptide sequences with their probability scores by MASCOT, whereas the other 297 peak lists were assigned to no sequence in the database searched. Of the 1837 peptide assignments, 232 were assigned at the top rank position to peptide sequences from the 14 subunits which compose the 20S proteasome complex. The remaining 1605 assignments were from other sequences.

The 1837 peptide assignments were analyzed with the ScoreRidge program. LC elution times for the peptide assignments were calculated using the modified retention coefficient values (Table 1). Figure 2a shows a scattered plot of the 1837 peptide assignments as the scan numbers *versus* calculated elution values. These peptide assignments,

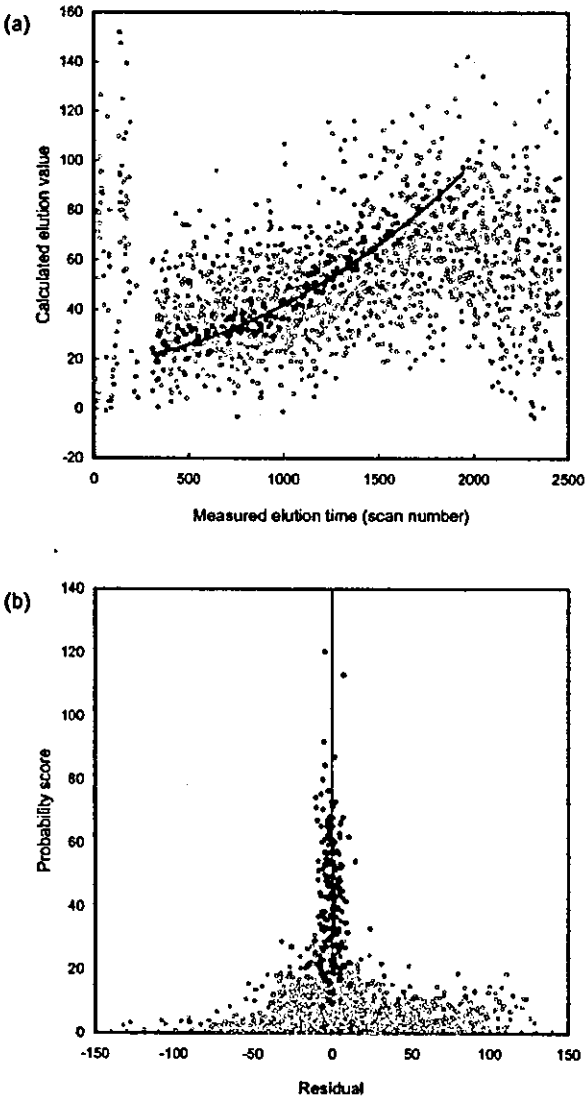

Figure 2. Distribution of 1837 peak lists assigned to proteasome peptide sequences (•) and other sequences (○). (a) Measured elution time *versus* calculated elution value. (b) Residual, prediction error value to the correlation curve shown in (a), *versus* peptide probability score.

probability scores, MS scan numbers and calculated elution values, were accumulated in order of height of the probability scores by a five unit interval. Figure 3 shows the cumulative distribution of probability scores for the 1837 peak lists. Numbers of peak lists within each probability score range are shown on top of the bar graphs. These accumulated data sets were used to develop regression equations for the correlation curves between the scan numbers and calculated elution values and to calculate the r-values for the correlation curves. In the program, the regression equations can be found at several approximation modes (including linear and polynomial ones), depending on the actual LC conditions. In this simulation, the correlation curve ($r = 0.938$) found by accumulation of peak lists with more than forty units of probability score were selected for the optimum PSEC equation. The correlation curve was described at quadratic polynomial approximation as shown in Fig. 2a. Essentially, most correct assignments indicated the tendency to concentrate nearby the correlation curve. Figure 2b shows another view for the 1837 peptide assignments: Residual, error prediction value to the correlation curve *versus* peptide probability scores.

To get an optimum prediction tolerance to accept the peptide assignments, numbers of peak lists within given tolerances were plotted by the absolute tolerance unit 2.5 (Fig. 4). With increasing tolerance, assignments to proteasome peptide sequences increased nonlinearly, and were nearly stable at tolerance units of more than 10.00, whereas assignments to other peptide sequences increased linearly in the residual range indicated. At the absolute tolerance value of 12.5, 96.1% (223/232) of the proteasome peptide assignments were within the tolerance. Other peptide assignments were limited to 34.8% (558/1605) within the tolerance. The absolute tolerance value was applied to the data analysis of HBV core-binding proteins. \

3.4 Identification of host proteins which interact with human HBV core protein

HBV core protein is one of the four gene products coded in the viral genome. To elucidate functions of the protein in relation to interaction with protein molecules expressed in host cells, the host components specifically bound to the core protein were purified using a conventional pull-down assay system. A major band with M_r 33 000 (p33) on a SDS-polyacrylamide gel was obtained. The p33 band was excised from the gel plate and subjected to tryptic digestion [8] in order to obtain a peptide mixture. The mixture was analyzed with LC-MS/MS. A total of 771 product ion spectra were obtained in a single analytical run. These product ion spectra were converted to 904 peak lists. Obtained ion peak lists were examined for assignment to tryptic peptide fragments in the human proteome database. Of 904 peak lists, 826 were assigned to peptide sequences with probability scores by MASCOT. These peptide assignments were subjected to the ScoreRidge program. The program was run according to the conditions optimized for the test data set of 1837 peptide assignments from the 20S proteasome (prediction tolerance ± 12.5). The correlation curve was described at quadratic polynomial approximation. Peak lists accepted under the conditions (245) were arranged for protein identification lists again using MASCOT.

Figure 5 shows the cumulative distribution of probability scores of 826 peptide assignments. Unlike for the 20S proteasome, the constant r-value region before a marked decreasing indicated relatively small r-values (0.6), probably because of a partially purified sample. Peptide assignments with probability scores higher than 40 were selected for the PSEC equation. The protein scores of the top 15 entries

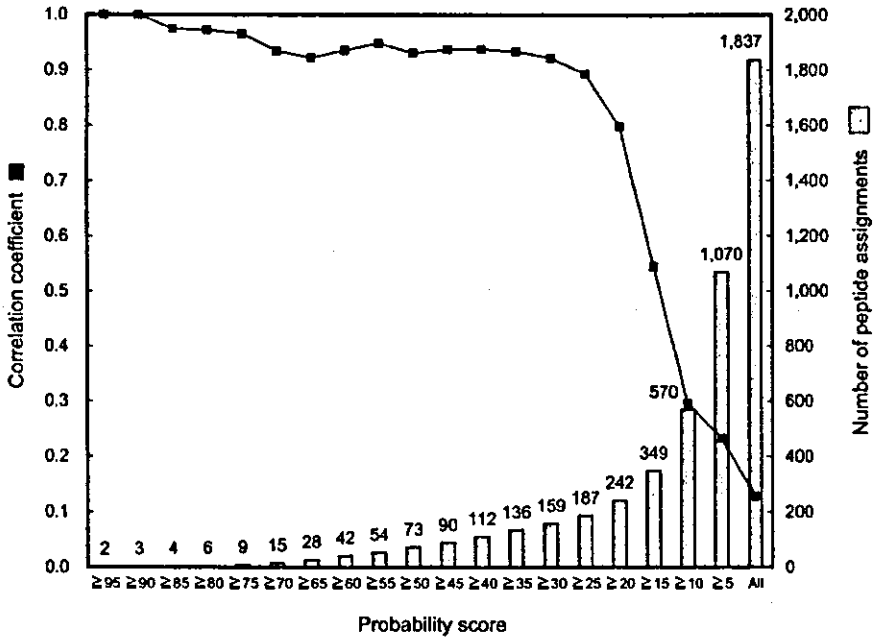

Figure 3. Cumulative distribution of probability scores of the 1837 peak lists with any peptide assignments obtained from analysis of the proteasome digest, and correlation coefficient values obtained from measured *versus* predicted elution of peak lists having probability scores within given ranges.

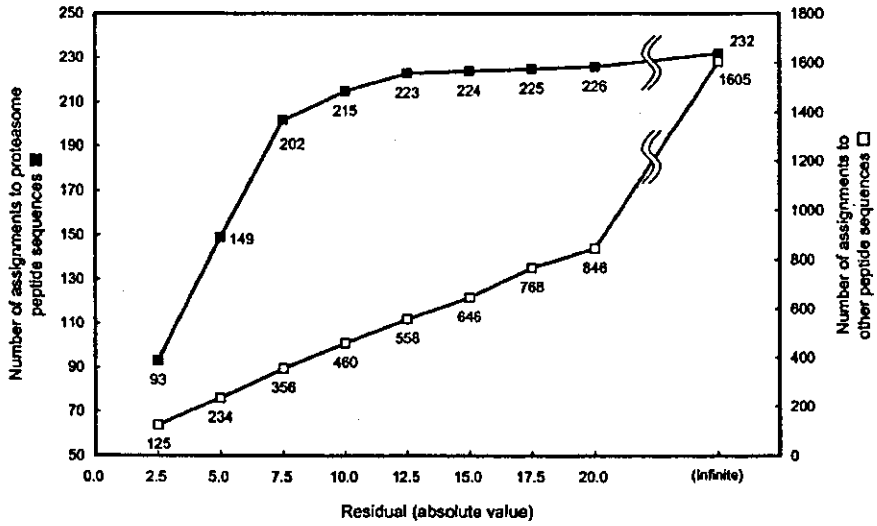

Figure 4. Cumulative distribution of numbers of peak lists, assigned to proteasome peptide sequences (■) and other sequences (□), on the given absolute residual values.

exceeded the probability threshold as significant protein matches. Four protein entries were selected as candidate proteins as shown in Table 2: carbonyl reductase (NADPH; EC 1.1.1.184) 1; complement component 1, Q subcomponent binding protein (C1qBP), mitochondrial; 40S ribosomal protein S2; 40S ribosomal protein S6. Only the molecular weight of sample protein (33,000) was considered in the final selection process. *N*-Acetylation of carbonyl reductase 1 was confirmed in the present study. Further investigation showed that C1qBP is a binding partner of HBV core protein under physiological conditions. The details and functional analyses, including *C1qBP* gene knock-down, will be discussed in a separate paper.

4 Concluding remarks

We have demonstrated in the present report that correct peptide assignments were selected at a sensitivity of more than 95% on the basis of peptide elution in RP LC and that the correlation curves from the higher-scored peptide assignments made it possible to flexibly use the ScoreRidge program. The program is applicable to different gradient rates and types without adjustments of the calculation parameters. When analyzing purified samples such as protein spots on 2-D gel plates, relatively small numbers of significant peptide assignments may be obtained from LC-MS/MS

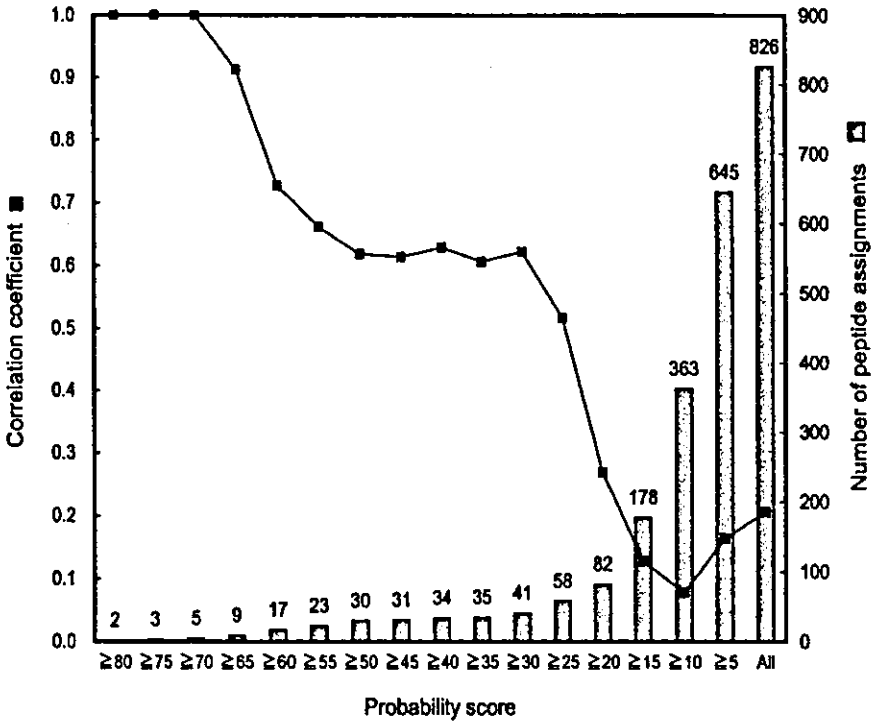

Figure 5. Cumulative distribution of probability scores of 826 peak lists with any peptide assignments obtained from analysis of the p33 protein band, and correlation coefficient values obtained from measured *versus* predicted elution of peak lists having probability scores within given ranges.

Table 2. Identification of host proteins interacting with HBV core protein

Protein name	Accesion no.	M_r[a]	Peptide assignment									
			No.	Charge	Observed mass (Da)	Expected mass (Da)	Calculated mass (Da)	Delta[b]	*MASCOT* score	Sequence[c]	Residual[d]	Peptide position[e]
Carbonyl reductase [NADPH]1	P16152	30,244	1	+2	400.4	798.8	798.0	0.8	57	(K)GIGLAIVR(D)	−1.33	15– 22
			2	+2	423.7	845.3	844.0	1.3	64	(K)IGVTVLSR(I)	−1.68	198–205
			3	+2	589.8	1177.7	1177.4	0.3	66	(R)LFSGDVVLTAR(D)	−5.26	27– 37
			4	+2	604.3	1206.5	1206.5	0.1	52	(R)VVNVSSIM*SVR(A)	−6.14	134–144
			5	+2	691.9	1381.7	1381.6	0.2	63	Acetyl-SSGIHVALVTGGNK(G)	−0.09	1– 14
			6	+2	800.7	1599.4	1599.8	−0.3	82	(R)FHQLDIDDLQSIR(A)	−1.27	58– 70
			7	+2	827.7	1653.4	1652.8	0.6	27	(R)GQAAVQQLQAEGLSPR(F)	2.55	42– 57
			8	+2	855.4	1708.8	1708.9	−0.1	58	(R)SETITEEELVGLM*NK(F)	−0.92	159–173
			9	+2	870.4	1738.8	1739.1	−0.3	55	(R)DVC*TELLPLIKPQGR(V)	0.80	119–133
			10	+2	890.7	1779.3	1780.0	−0.7	82	(K)EYGGLDVLVNNAGIAFK(V)	−2.40	79– 95
			11	+3	601.1	1800.1	1800.1	0.1	48	(K)VADPTPFHIQAEVTM*K(T)	−3.97	96–111
			12	+2	955.6	1909.2	1908.2	1.0	11	(R)KEYGGLDVLVNNAGIAFK(V)	−2.29	78– 95
			13	+3	671.6	2011.8	2012.3	−0.5	22	(K)FRSETITEEELVGLM*NK(F)	−8.10	157–173
			14	+2	1007.1	2012.1	2012.3	−0.1	18	(K)FRSETITEEELVGLM*NK(F)	−8.01	157–173
Complement component 1, Q subcomponent binding protein, mitochondrial	Q07021	31,362	15	+1	480.4	479.4	479.6	−0.2	18	(K)SFVK(S)	7.39	277–280
			16	+2	642.8	1283.7	1283.4	0.2	62	(K)AFVDFLSDEIKE(E)	1.27	81– 91
			17	+2	757.6	1513.1	1513.6	−0.4	68	(R)EVSFQSTGESEWK(D)	8.88	208–220
			18	+2	819.9	1637.7	1637.8	0.0	71	(K)M*SGGWELELNGTEAK(L)	3.21	105–119
			19	+2	849.9	1697.7	1697.9	−0.2	69	(K)AFVDFLSDEIKEER(K)	0.89	81– 94
			20	+3	567.5	1699.5	1697.9	1.7	65	(K)AFVDFLSDEIKEER(K)	0.99	81– 94
			21	+2	1432.7	2863.5	2863.1	0.4	41	(K)ITVTFNINNSIPPTFDGEEEP SQGQK(V)	−0.38	129–154

Table 2. Continued

Protein name	Accesion no.	M_r[a]	Peptide assignment									
			No.	Charge	Observed mass (Da)	Expected mass (Da)	Calculated mass (Da)	Delta[b]	*Mascot* score	Sequence[c]	Residual[d]	Peptide position[e]
40S ribosomal protein S2	P15880	31,324	22	+2	777.4	1552.7	1551.8	0.9	60	(K)SLEEIYLFSLPIK(E)	−11.07	77– 89
			23	+2	1077.5	2152.9	2153.5	−0.6	17	(K)ESEIIDFFLGASLKDEVLK(I)	−10.12	90–108
40S rivosomal protein S6	P10660	28,681	24	+2	643.4	1284.7	1284.4	0.3	31	(K)DIPGLTDTTVPR(R)	7.42	120–131

a) Calculated value from the protein sequence, not taking into account any annotated PTM
b) Subtraction of calculated mass value from expected mass value
c) C* and M* stand for carboxyamidomethyl cysteine residue and methionine sulfoxide residue, respectively
d) Subtraction of elution value calculated from amino acid sequence from elution value estimated from correlation curve
e) The numbers of *N*-terminal and *C*-terminal amino acid residues of peptides in the protein sequence

data. In such cases, the samples may be added with known peptide mixtures before analysis in order to increase reference points composing the correlation curve.

The ScoreRidge program has the ability to complement product ion search engines which express probability of peptide matching by a numerical value. Once integrated into conventional search engines, ScoreRidge is applicable to many kinds of studies including large-scale proteome analysis using multidimensional LC-MS/MS [2, 3, 18]. Qualitative analyses such as cataloging thousands of protein names, selection of positive peptide assignments and exclusion of most negative assignments will result in reduced numbers of false negative and false positive protein entries in the final identification lists. More importantly, in the case of ion intensity-based quantitative analyses using various chemical or metabolical tagging methods, peptide assignments from small numbers of precursor ions significantly changed in their relative intensities may be pursued to identify up- or down-regulated proteins by the ScoreRidge complement. Recent advances in LC-MS/MS instruments in terms of stability and robustness have made it possible to compare quantitatively multiple mass chromatograms. Also in this case, the program was effective in confirming protein identities with significant differences in peptide ion intensities (manuscripts in preparation). Use of the present simple and practical program will reduce laborious manual interpretation of search results and consequently improve both reliability and throughput of LC-MS/MS data analysis in the protein research fields and especially in proteomics.

The authors thank Dr. A. Ogiwara, (Clinical Proteome Center, Tokyo Medical University), for his useful suggestions. This work was supported in part by Grants-in-Aid for scientific research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

5 References

[1] Link, A. J. (Ed.), *2-D Proteome Analysis Protocols*, Humana Press, Totowa, NJ, USA 1999, pp

[2] Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., Gygi, S. P., *J. Proteome Res.* 2003, *2*, 43–50.

[3] Shen, Y., Jacobs, J. M., Camp II, D. G., Fang, R. *et al.*, *Anal. Chem.* 2004, *76*, 1134–1144.

[4] Meek, J. L., *Proc. Natl. Acad. Sci. USA* 1980, *77*, 1632–1636.

[5] Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A. *et al.*, *Anal Chem.* 2003, *75*, 1039–1048.

[6] Kanai, F., Marignani, P. A., Sarbassova, D., Yagi, R. *et al.*, *EMBO J.* 2000, *19*, 6778–6791.

[7] Kuroki, K., Cheung, R., Marion, P. L., Ganem, D., *J. Virol*, 1994, *68*, 2091–2096.

[8] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal Chem.* 1996, *68*, 850–858.

[9] Pappin, D. J. C., Hojrup, P., Bleasby, A. J., *Curr. Biol.* 1993,*3*, 327–332.

[10] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[11] Creasy, D. M., Cottrell, J. S., *Proteomics* 2002, *2*, 1426–1434.

[12] Arnott, D. P., Gawinowicz, M., Grant, R. A., Lane, W. S. *et al.*, *J. Biomol. Tech.* 2002, *13*, 179–186.

[13] Eng, J. K., McCormack, A. L., Yates, J. R. 3rd, *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[14] MacCoss, M. J., Wu, C. C., Yates, J. R. 3rd, *Anal Chem.* 2002,*74*, 5593–5599.

[15] http://www.chem.agilent.com.

[16] Jensen, O. N., Wilm, M., Shevchenko, A., Mann, M., in: Link, A. J. (Ed.), *2-D Proteome Analysis Protocols*, Humana Press, Totowa, NJ, USA 1999, pp. 513–530.

[17] Polevada, B., Sherman, F., *J. Mol. Biol.* 2003, *325*, 595–622.

[18] Link, A. J., *Trends Biotechnol.* 2002, *20*, S8–13.

research articles Journal of proteome •research

Multidimensional Protein Profiling Technology and Its Application to Human Plasma Proteome

Kiyonaga Fujii,[†,§] Tomoyo Nakano,[‡] Takeshi Kawamura,[†] Fumihiko Usui,[‡] Yasuhiko Bando,[‡] Rong Wang,[§] and Toshihide Nishimura*[,†]

Clinical Proteome Center, Tokyo Medical University, 2-6-1, Nishi-shinjuku Shinjuku-ku, Tokyo 163-0203, Japan, AMR Incorporated, 2-13-18, Nakane Meguro-ku, Tokyo 152-0031, Japan, and Department of Human Genetics, Mount Sinai School of Medicine, 1425 Madison Avenue Box 14–52, New York, New York 10029-6574

Received November 7, 2003

In clinical and diagnostic proteomics, it is essential to develop a comprehensive and robust system for proteome analysis. Although multidimensional liquid chromatography/tandem mass spectrometry (LC/MS/MS) systems have been recently developed as powerful tools especially for identification of protein complexes, these systems still some drawbacks in their application to clinical research that requires an analysis of a large number of human samples. Therefore, in this study, we have constructed a technically simple and high throughput protein profiling system comprising a two-dimensional (2D)-LC/MS/MS system which integrates both a strong cation exchange (SCX) chromatography and a μLC/MS/MS system with micro-flowing reversed-phase chromatography. Using the μLC/MS/MS system as the second dimensional chromatography, SCX separation has been optimized as an off-line first dimensional peptide fractionation. To evaluate the performance of the constructed 2D-LC/MS/MS system, the results of detection and identification of proteins were compared using digests mixtures of 6 authentic proteins with those obtained using one-dimensional μLC/MS/MS system. The number of peptide fragments detected and the coverage of protein sequence were found to be more than double through the use of our newly built 2D-LC/MS/MS system. Furthermore, this multidimensional protein profiling system has been applied to plasma proteome in order to examine its feasibility for clinical proteomics. The experimental results revealed the identification of 174 proteins from one serum sample depleted HSA and IgG which corresponds to only 1 μL of plasma, and the total analysis run time was less than half a day, indicating a fairly high possibility of practicing clinical proteomics in a high throughput manner.

Keywords: proteomics • human plasma • multidimensional chromatography • μLC/MS/MS

1. Introduction

Proteomics provides a powerful tool to gain deep insights into disease mechanisms in which proteins play a major role. It can also investigate structures and functions of protein complexes working as molecular micro-machines expressed spatiotemporally through protein–protein interactions and signal transduction by post-translational modification.[1] Protein profiling has often been performed by the "classical" one- or two-dimensional sodium dodecyl sulfate polyacryamide gel electrophoresis (1D- or 2D-PAGE) based on the densitometric quantification of proteins visualized with using dyes on gel. After in-gel enzymatic digestion of the subject protein sports, the resulting peptides are subjected to matrix-assisted laser desorption ionization (MALDI) or electrospray ionization (ESI) mass spectrometry. Queries concerning mass-spectrometric data-sets are provided to the protein/nucleic acid databases to identify proteins and their modifications.[1–4]

Recently, the series of procedures using 2D-PAGE are routinely practiced in most proteomics laboratories, because 2D-PAGE can separate protein isoforms in high resolution and still be powerful in profiling post-translational modifications. However, gel-based proteomics faces various shortcomings, such as being time-consuming, creating a heavy workload, lack of sufficient reproducibility, poor recovery in in-gel proteolysis, insolubility of hydrophobic proteins, and difficulty in detection and separation of low-abundance proteins.[2–5]

Due to these shortcomings in 2D-PAGE, especially in proteome analysis on an industrial or clinical scale, we investigated new methods of high throughput analysis instead of gel-based protein separation technologies. We had previously employed the μLC/MS/MS system with reversed-phase chromatography at a micro-flow rate and nanoESI (NSI) interface to find lung cancer biomarkers as well as to analyze the apoptopic mechanisms, following the preparation of peptide mixtures by means of in-gel digestion of the 24 fractionations of 1D-PAGE.[6,7] Since

* To whom correspondence should be addressed. Tel.: +81-3-5321-6623. Fax: +81-3-5321-6624. E-mail: nisimura@tokyo-med.ac.jp.
† Clinical Proteome Center, Tokyo Medical University.
‡ AMR Incorporated.
§ Department of Human Genetics, Mount Sinai School of Medicine.

Published on Web 06/09/2004
10.1021/pr030007q CCC: $27.50

our protein profiling μLC/MS/MS system involves desalting and concentration systems by trap-cartridge for the injected samples before loading into reversed-phase column, it can be useful in improving a high throughput system and in connecting with other various separation systems. Moreover, the reversed-phase system is advantageous in separating various peptides with relatively high resolution and results in protein identification with high sequence coverage.

More recently, multidimensional LC/MS/MS systems have been developed for powerful and comprehensive protein identification in protein complexes and have achieved resolution power compatible to 2D-PAGE, however, these needs to be developed further.[2-5] Washburn et al. have demonstrated the profile analysis of the whole yeast proteome using their multidimensional protein identification technology, which involves 2D chromatography for peptide mixtures and utilizes an orthogonal biphasic capillary column packed with both strong cation exchange (SCX) and reversed-phase materials, in tandem.[8,9] However, the use of biphasic capillary column leads to difficulties in terms of ease in usage, mechanical stability, and industrial and clinical applicability. Various groups have attempted to further refine and optimize the construction of multidimensional chromatography coupled with mass spectrometry specifically for large-scale protein profiling.[3,5,10-12]

In clinical and diagnostic applications where highly complex protein mixtures are subjected to proteome analysis, it is essential to establish a global platform for large-scale as well as a simple, robust, and high protein profiling methodology. Such a method is needed because a huge number of human samples of tissues/biological fluids would be subjected in routine and reproducible quantitative analysis for protein expression in order to discover a protein that is significantly associated with disease status. Hence, for the routine clinical use, we tried to construct a technically simple and high throughput 2D-LC/MS/MS system as a multidimensional protein profiling technology in which SCX chromatography as a first dimension was connected to our established μLC/MS/MS system.

This report describes the construction and evaluation of our comprehensive 2D-LC/MS/MS system and its application to proteome analysis of human plasma.

2. Experiments

2.1. Materials. HPLC-grade acetonitrile, formic acid, and trifluoroacetic acid (TFA) were purchased from Wako Pure Chemical Industries, Ltd. (Osaka, Japan). The water used was Milli-Q grade (Millipore, Bedford, MA). Six bovine proteins (β-lactoglobulin, glutamic dehydrogenase, bovine serum albumin, apotransferrin, lactoperoxidase, and catalase), human plasma, ammonium formate, ammonium bicarbonate (ABC), and iodoacetamide (IAM) were purchased from Sigma (St. Louis, MO). Tris[2-carboxyethyl]phosphine (TCEP) was obtained from Pierce (Rockford, IL). Sequencing grade-modified trypsin was a product of Promega (Madison, WI).

2.2. Sample Preparation of Digests Consisting of 6-Mixed Standard Proteins. Digest samples consisting of 6-mixed standard proteins (6-protein digests) were prepared by mixing six individual standard cow proteins, β-lactoglobulin, glutamic dehydrogenase, bovine serum albumin, apotransferrin, lactoperoxidase and catalase after in-solution digestion. 1 nmol of each protein was diluted with 230 μL of 100 mM ABC. Then 12.5 μL of 10 mM TCEP was added for reduction and the mixing solution was kept at 37 °C for 45 min. Next, 12.5 μL of 50 mM IAM was added and the mixing solution was treated for alkylation at 24 °C for 1 h in darkness. The resultant samples were digested by the tryptic protease: protein ratio of 1:50 (w/w), and all the resultant 250 μl. solutions were incubated at 37 °C for 15 h in darkness. All of these reactions utilized an Eppendorf thermomixer R (Brinkmann, Westbury, NY) for a 1.5 mL microtube and interval-mixed (10 s) at 850 rpm. Each 5 μL aliquot of the digest peptide samples from β-lactoglobulin, glutamic dehydrogenase, bovine serum albumin, and apotransferrin, as well as both 2.5 μL aliquots of those from lactoperoxidase and catalase were mixed in a 200 μL auto sampler tube, methylpentene polymer (TPX) tube (Hitech Inc. Tokyo, Japan) and filled up to 200 μL with 2% acetonitrile (aq.) containing 0.1% TFA for the 1D-μLC/MS/MS analysis (each 50−100 fmol/μL). For the 2D-LC/MS/MS analysis, volumes corresponding to 10 times that of the original aliquots (i.e., 25−50 μL, digested samples corresponding to 100−200 pmol) were mixed in the same sample mixing ratio and diluted with 250 μL of 2% acetonitrile (aq.) containing 0.005% TFA (aq.) after adjusting the pH level to about 3 with 500 μL of 0.5% TFA (aq.).

2.3. Sample Preparation of the Digested Human Plasma Protein Mixture. Human serum albumin (HSA) and immunoglobulin G (IgG) in 500 μL of human plasma (31 mg protein/mL) were depleted by affinity adsorption chromatography using Bio-Rad's Affi-Gel Blue Gel and protein A column (Bio-Rad Laboratories, CA), respectively. The resulting HSA- and IgG-depleted human plasma (AID-HP) sample was finally brought to a concentration of 3.5 mg/mL in 50 mM of ABC (details not shown). Then, 250 μL of the AID-HP sample was diluted with 390 μL of 50 mM ABC and 250 μL of acetonitrile. For reduction, 50 μL of 10 mM TCEP was added and the mixing solution was kept at 37 °C for 45 min. Then, 50 μL of 50 mM IAM was added, and the mixing solution was treated for alkylation at 24 °C for 1 h in darkness. In digestion, 10 μg of trypsin was added and total 1.0 mL of the resulting solution was incubated at 37 °C for 18 h in darkness. All these reactions were carried out using the Eppendorf thermomixer R for 1.5 mL microtubes (Sorenson BioScience Inc. UT) and mixing was carried out at 850 rpm for periods of 10 s at intervals of 10 s. To obtain equivalent amounts to the 50 μg proteins, 57 μL of the digested AID-HP sample was diluted to 200 μL with 2% acetonitrile (aq.) containing 0.1% TFA in a TPX auto sampler tube for the 1D-LC/MS/MS analysis. For the 2D-LC/MS/MS analysis, the same 57 μL was diluted with 86 μL of 2% acetonitrile (aq.) containing 0.005% TFA (aq.) after adjusting the pH level to about 3 with 57 μL of 1% TFA (aq.).

2.4. SCX Separation. Separation of the resultant peptide fragments was performed on a SCX MicroBullet cartridge (12 μm, 300 Å, 25 × 0.5 — 2.0 mm i.d., Michrom BioResources, Inc., Auburn, CA) using a KDS100 syringe pump (KD Scientific Inc. PA) as a solvent delivery system. The solvent system was composed of mobile phase C, 0.005% TFA (aq.) and mobile phase D, 1 M ammonium formate (aq.) adjusted to pH 3.2 with formic acid, and both of them contained with 2% acetonitrile. Elution solvents (25, 50, 100, 150, 200, 250, 300, and 500 mM) were respectively prepared by mixing the mobile phase C and D, and peptides were eluted stepwise at 100 μL/min by 200 μL of each elution solvent. Each stepwise effluent was collected manually in a 200 μL HPLC sample tube. After vortex, the resulting sample solution was directly subjected to reversed-phase chromatography followed by LCQ mass spectrometry.

2.5. Reversed-Phase Separation and NSI−MS/MS System. The μ-HPLC/NSI−MS/MS system was composed of a MAGIC

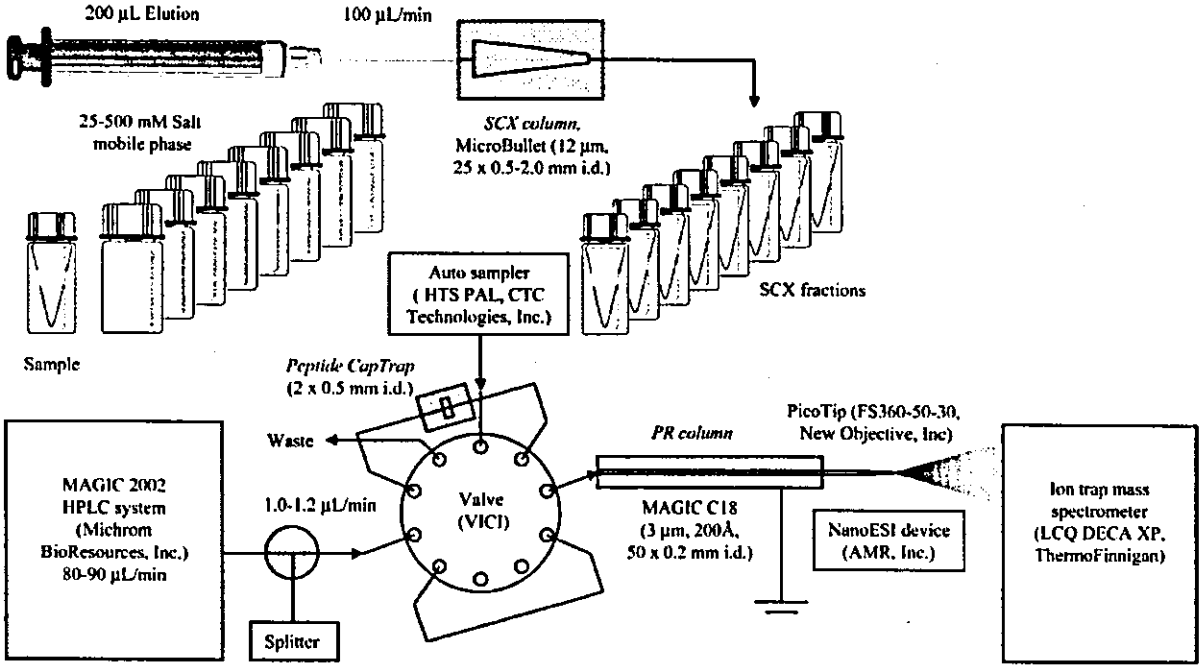

Figure 1. Schematic drawing of the off-line 2D-µLC/NSI-MS/MS system used in this report (not to scale).

2002 dual solvent delivery system (Michrom) for HPLC, a HTS PAL auto sampler (CTC Analytics, Zwingen, Switzerland) and a Finnigan LCQ Deca XP ion trap mass spectrometer (Thermo Electron, San Jose, CA) equipped with a nano-electrospray ion (NSI) source (AMR Inc, Tokyo, Japan). Digested samples and SCX fractions were automatically injected into a peptide CapTrap cartridge (2.0 × 0.5 mm i.d., Michrom) on an injector valve for concentration and desalting. After desalting with 0.1% TFA (aq.) containing 2% acetonitrile, the sample was loaded into a reversed-phase column, MAGIC C18 (3 µm, 200 Å, 50 × 0.2 mm i.d., Michrom) for separation. The solutions of 2% and 90% acetnitrile (aq.) were used as mobile phase A and B, respectively, and both contained with 0.1% formic acid. The gradient conditions in the chromatographic run were set up as follows: program 1, B 5% (0 min) → 40% (35 min) → 95% (40 min); program 2, B 5% (0 min) → 40% (70 min) → 95% (80 min). Effluent from the HPLC at a flow rate of 80 µL/min was split using a MAGIC variable splitter (R2 position, Michrom), and the effluent at 1.0−1.2 µL/min was introduced into the mass spectrometer by the NSI interface via an injector valve with a CapTrap cartridge and the column. The NSI needle (PicoTip FS360−50−30, New Objective Inc., Woburn, MA) attached directly to the reversed-phase column was used as the NSI interface and the voltage was 1.7 kV, whereas the capillary was heated to 250 °C. No sheath or auxiliary gas was used. Furthermore, the mass spectrometer was operated in a data-dependent acquisition mode in which the MS acquisition with a mass range of m/z 450−2000 was automatically switched to MS/MS acquisition under the automated control of Xcalibur software. The most intense ion of the full MS scan was selected as the parent ion and subjected to MS/MS scan with an isolation width of m/z 2.0, and the activation amplitude parameter was set at 35%. For the samples of human plasma proteins, the full MS scan was acquired followed by the three successive MS/MS scans of the three most intense precursor ions detected in the full MS scan. The trapping time was 200 ms under the auto gain control mode. Data acquisition was carried out using the dynamic mass-exclusion windows that had an exclusion of 1.0 min duration and exclusion mass width of ±1.5 Da.

2.6. Database Searches. All MS/MS data were investigated using the MASCOT search engine (Matrix Science, Ltd., London, UK) against the Swiss-Prot database. The data acquired for G-protein digests were investigated against the other mammalian subsets of the sequences. For the sample of human plasma proteins. The MS/MS data were investigated only against the Homo sapiens (human) subset of the sequences. The database searches were performed allowing for fixed modification on cysteine residue (carbamidomethylation, +57 Da) and variable modification on methionine residue (oxidation, +16 Da), peptide mass tolerance at ±2.0 Da, and fragment mass tolerance at ±0.8 Da.

3. Results and Discussion

3.1. Configuration of 2D-LC/MS/MS System. Figure 1 schematically illustrates our 2D-LC/MS/MS protein profiling system. The µLC/MS/MS system with reversed-phase separation (1D-RP analysis), which corresponds to the second dimensional separation, is composed of a micro-flow LC system, a versatile auto-sampler equipped with an injector valve, and a LCQ ion-trap mass spectrometer with an NSI stage. A flow rate of 1.0−1.2 µL/min via an injector valve and reversed-phase column is obtained by splitting the elution solvent flow rate of about 50−100 µL/min from the dual pumps using a variable splitter. The concentration and the desalting of peptide samples are achieved by loading them onto a trap cartridge of 0.5 µL volume prior to column separation. After washing the trap cartridge using the auto-sampler, the sample is injected into the reversed-phase column. Our NSI interface has a spray needle designed to be attached directly to the column end, so that the sample is electro-sprayed with minimum diffusion of the separating sample due to suppression of the dead volume that usually exists in conventional tubing after its elution from

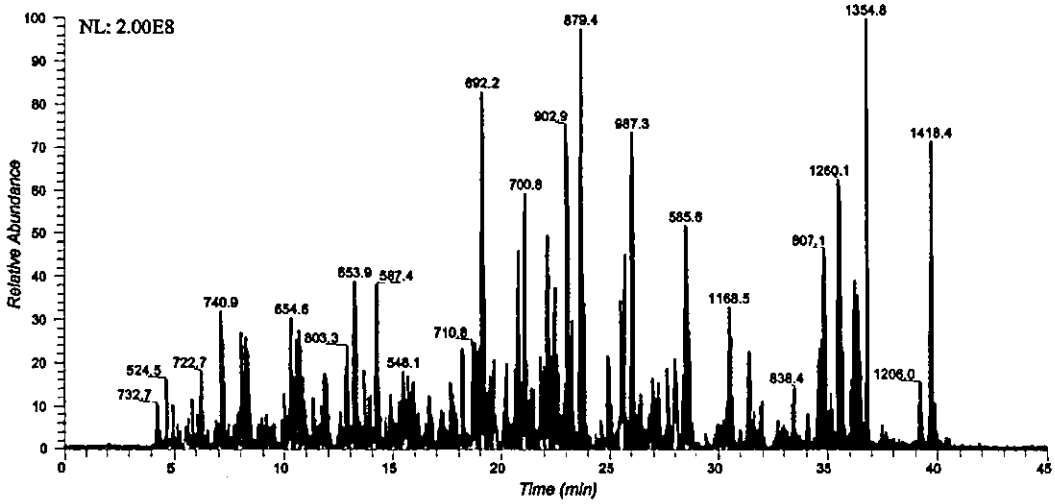

Figure 2. Base peak chromatogram of digests consisting of 6-mixed standard proteins analyzed by the 1D-μLC/MS/MS system.

the column. MS/MS spectra are data-dependently accumulated following MS measurement for precursor peptide ions with dynamic mass exclusion windows in the ion-trap mass spectrometer. An LC analytical running time of 20 min is sufficient under a linear gradient of 3% B/min to analyze peptide samples obtained from one or a few proteins such as in-gel digestion samples; however, a longer running time is expected for additional complex protein digests in order to obtain more MS/MS spectra. Thus, for the second dimensional chromatography in the 2D-LC/MS/MS system, 40 min was chosen as the running time under the gradient condition of 1% B/min, taking into account the peak width (0.18−0.24 min at 50% peak height) and the acquisition efficiency of the MS/MS spectra. Alternatively, a running time of 80 min was also tested under the gradient condition of 0.5% B/min (peak width, 0.23−0.31 min at 50% peak height). Figure 2 shows the mass chromatogram obtained for 6-protein digests (250−500 fmol, respectively). The detection limit for protein identification was around 5−10 fmol of protein digests.

To construct the 2D-LC/MS/MS system (2D-LC analysis), we utilized SCX separation as the first dimensional separation. SCX chromatography is performed by 8−9 stepwise fractionations using elution solutions of different salt concentrations. SCX fractionated samples are subjected to 1D-RP analysis for second dimensional separation without further sample preparation. This protein profiling system using multidimensional LC/MS/MS can analyze proteomes at a throughput rate of two samples per day. An SCX column with 100 μg capacity was developed, taking into consideration the capacity of the trap cartridge (maximum 2 μg), the elution volume and the number of SCX fractionations. Additionally, a bullet type SCX column was used to improve the resolution during separation. Ammonium formate was chosen as volatile buffer salt in the SCX mobile phase, avoiding nonvolatile salts such as NaCl, which is compatible with the solvent system of formic acid used in the second dimensional 1D-RP analysis.

As a preliminary experiment, elution volume and salt concentration for the SCX separation were examined, and it was suggested that the elution of a fraction can be achieved at over 4 times the column volume (>100 μL) and all samples absorbed with SCX can be eluted with the 500 mM salt concentration (data not shown). Therefore, it was determined that the stepwise elution would be carried out with 9 fractionations using 25, 50, 100, 150, 200, 250, 300, and 500 mM salt concentration with each 200 μL elution volume together with the pass-through fractions occurring at the time of sample loading.

3.2. Evaluation of the Constructed 2D-LC/MS/MS System. To evaluate the performance of our system, results obtained for both detection and protein identification were compared with those obtained by 1D-RP analysis. Six standard proteins with a wide range of molecular weights (18−128 kDa) were used, and 1D-RP and 2D-LC analyses were carried out after their in-solution digestion.[13] The 6-protein digests (total amount of about 65 μg containing each 100−200 pmol proteins) were applied to the SCX column and fractionated by 200 μL stepwise elution. Each fractionated eluent was collected into the 200 μL autosampler tube, and the sample solution of 1 μL (corresponding to 250−500 fmol of each protein) was directly injected into the μLC/MS/MS system (1D-RP analysis) for second dimensional chromatography. It was found that almost all peptides were fractionated into respective fractions of different salt concentrations at a high yield while a few peptides were the fractions of a dispersed over two fractions as the carry-over. On the other hand, 1D-RP analysis with no SCX separation was performed for the sample of the same amount (250−500 fmol) as in the 2D-LC analysis. The base-peak chromatograms of 6-protein digests obtained by 1D-RP analysis are shown in Figure 2.

The protein identification for the resulting MS and MS/MS spectra was carried out using the MASCOT database search software.[14] The number of peptide fragments identified as the six standard proteins with scores higher than 20 for MS/MS spectra are summarized in Figure 3. Under the analytical and database search thresholds, for MS/MS score higher than 20 and 30 on the MASCOT database search results, their statistical confidence were 76% and 97%, respectively (data not shown). In the experiment involving the 1D-RP analysis, about 80 peptide fragments were detected and identified, giving about 35% coverage in each protein whereas about 70% coverage was obtained in the 1D-RP analysis of the individual protein digests. On the other hand, in the 2D-LC analysis, about 35 peptide fragments were detected and identified on an average at each SCX fraction except for the pass-through fraction in which no

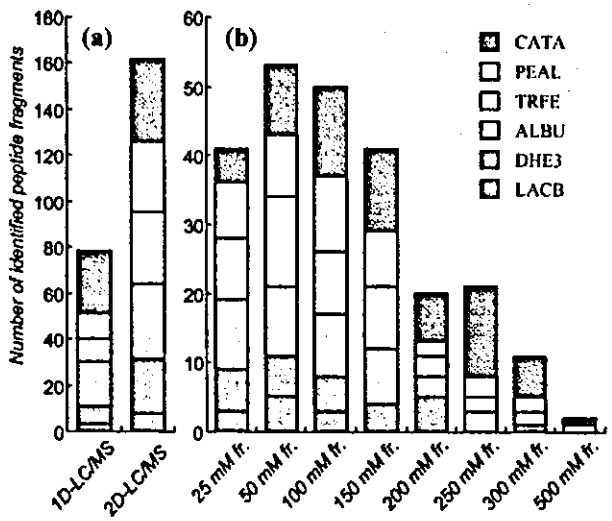

Figure 3. Database search results of the digests consisting of 6-mixed standard proteins analyzed by the 2D-LC/MS/MS system using Mascot Protein Identification. CATA, catalase; PEAL, lactoperoxidase; TRFE, apotransferrin; ALBU, bovine serum albumin; DHE3, glutamic dehydrogenase; LACB, β-lactoglobulin.

peptides were eluted (Figure 3). The total number of nonduplicated peptide fragments detected was found to be twice that of 1D-RP analysis. The sequence coverage in each protein increased twofold, indicating that almost all peptide fragments in the mixed protein sample could be detected and identified by 2D-LC analysis. Thus, it was conclusively demonstrated that the newly constructed 2D-LC/MS/MS system constructed by us has a capability of high resolution peptide mapping, which is sufficient to identify protein components in a highly complex protein mixture.

Table 1. Database Search Results of the Digested Human Plasma Sample Analyzed by the 2D-μLC/MS/MS System Using Mascot Protein Identification

	total[a]		> 20[a]		> 30[a]	
	query[b]	protein[c]	peptide[d]	protein[c]	peptide[d]	protein[c]
1D-RP analysis	3038	1020	606	116	431	51
2D-LC analysis	20760	2438	2935	713	1956	174
pass-through fr.	1608	254	224	87	143	40
25 mM fr.	1880	466	348	123	244	60
50 mM fr.	2501	1141	564	204	349	74
100 mM fr.	2569	760	470	136	316	61
150 mM fr.	3011	915	501	143	345	66
200 mM fr.	3077	611	416	91	294	48
250 mM fr.	2822	483	264	70	177	31
500 mM fr.	3292	362	148	61	88	21

[a] Score of MS/MS data: total, containing all score; > 20, score 20 and over; > 30, score 30 and over. [b] The numbers of dta files originated from MS/MS data. [c] The numbers of hit proteins. [d] The numbers of hit peptide fragments against the database.

3.3. Application of the 2D-LC/MS/MS System to Human Plasma Proteomics. To confirm the usefulness of the 2D-LC/

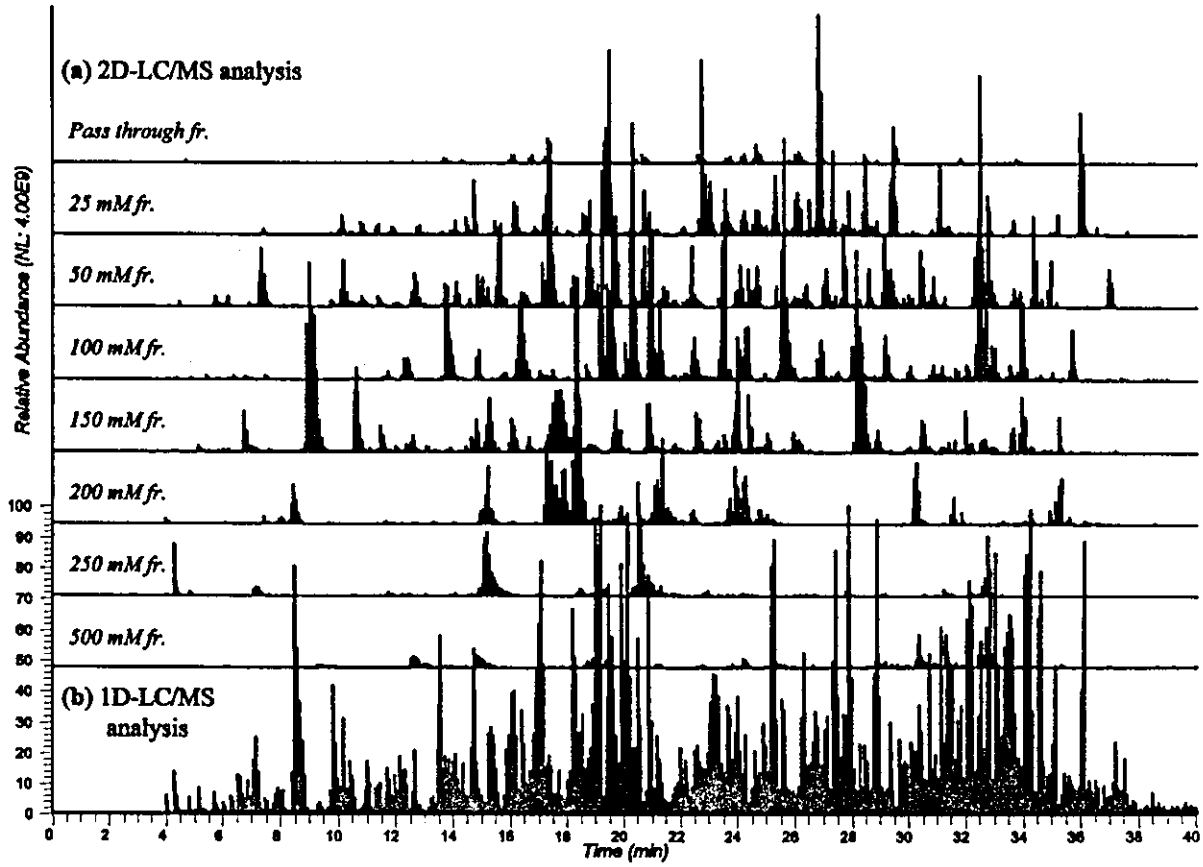

Figure 4. Base-peak chromatograms of the digested human plasma sample analyzed by the (a) 2D-LC/MS/MS systems and (b) 1D-LC/MS/MS systems.